DOCUMENT RESUME

ED 172 085 CG 013 527

AUTHOR Downey, Ronald G.; Saal, Frank E.
TITLE Evaluating Human Judgement Techniques.
PUB DATE 78
NOTE 20p.; Paper presented at the Annual Convention of the American Psychological Association (Toronto, Ontario, Canada, August, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adults; *Behavior Patterns; *Behavior Rating Scales; *Evaluation Criteria; *Evaluation Methods; *Performance Criteria; *Psychometrics; Reactive Behavior; Testing Problems
IDENTIFIERS *Judgment

ABSTRACT

The development of some psychometric indices of performance rating quality based in judgement is traced. The conceptual and analytical confusion that characterizes contemporary research in this area is described. The operational definitions of several criteria of rating quality, including leniency and severity of the rating effect, halo effect, central tendency, and reliability are discussed. A set of positively-toned criteria for evaluating the quality of rating data as well as the suitability of some common research designs for generating statistical indices is presented. (Author/HLM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED172085

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald G. Downey

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

EVALUATING HUMAN JUDGEMENT TECHNIQUES

Ronald G. Downey and Frank E. Saal

Kansas State University

CG013527

Paper prepared for 1978 American Psychological Association Meeting in Toronto, Canada.

Evaluating Human Judgment Techniques

Background

There are few areas of our lives that have not been affected by quantified human judgments in the form of ratings. The experimental and theoretical history of ratings dates, as far as we can determine, to the early 1920's, and is important from two perspectives. First, it documents the beginnings of many of the issues with which we are still trying to grapple today; second, it clearly establishes many of the confusions still plaguing the research and applied work in the area.

Although Wells (1909) and Webb (1915) seem to have identified the process earlier, one of the earliest landmarks is Thorndike's (1920) research on halo, or the rater's tendency to think of a person in general as good or inferior, resulting in high intercorrelations between dimensions across ratees. This work was followed very closely by the prodigious work of Rugg (1921a, 1921b, 1922a, 1922b), dealing with personnel ratings in the Army and paralleling the Alpha Beta work in group intelligence testing. Rugg's stated purpose was to answer the question: "Can human character be rated on point scales accurately enough for practical uses in education?" The primary evaluation technique used was interrater agreement. Rugg (1922a, p. 39) concluded that accurate ratings can be obtained by averaging several judgements, and felt that the failure of a single judge's (rater's) estimate was due to "general attitudes toward our associates and subordinates" (halo). Kingsbury (1922) seems to be the first reference that lists the major criteria used today for evaluating ratings; (1) halo (Thorndike, 1920); (2) high-low raters (leniency); and (3) fear of making distinctions (central tendency). No attempt is made to quantify these errors, however. Kneeland (1929) seems to have coined the phrase "lenient" to express the tendency of raters to "rate well above the mid-point

of the scales used, as indicated by average ratings" (averaged over all ratees). Ford (1931) adopted the term lenient and added the term "severe" to describe raters who concentrate on the low end of the scale. He also introduced a method of correcting the scores for these tendencies. Newcomb (1931, p. 288) identified a new type of error, logical error, in which "close relation between the intra-trait behaviors (correlations) . . . be presumed to spring from logical presuppositions in the minds of the raters, . . ." The logical error bears a striking resemblance to the contemporary notion of implicit personality theory (Schneider, 1973). Murray (1938) identified another error, contrast error, which is the tendency of raters to compare ratees against themselves. Finally, Stockford and Bissell (1949) found that the degree of correlation between traits varied as a function of the physical distance between traits on the rating form, and labeled this "proximity" error. Although others (mentioned later) have appeared periodically, there "errors" are most commonly used to assess the value or quality of a rating procedure.

The purpose of this brief historical review was to establish the demarcation points for current work in the area, in an attempt to facilitate our understanding of these errors. We chose, somewhat at random, several classic and more recent textbooks in the industrial and psychometric areas, and catalogued their mention of the reviewed types of rating errors, definitions given, and methods for quantifying these errors. The results are shown in Table 1, which highlights several important findings. First, halo, central tendency and leniency were common to all the books. Second, while definitions were generally in agreement, only the methods for quantifying halo showed appreciable consistency; the methods for quantifying central tendency and leniency were more variable. Third, with the exceptions of Guilford (1954) and Smith (1976), halo was the only error

thought to contribute to trait intercorrelations. Finally, those sources that did not suggest methods for quantifying the errors generally also failed to refer to sources containing quantification strategies.

Although not discussed above, reliability has also been utilized as a criterion of rating quality, and has usually been conceptualized as either inter-rater agreement (see Ebel, 1951) or intra-rater agreement over time. Methods for estimating an index of agreement can differ according to the situation (see Ebel, 1951 for a more detailed explanation).

Another approach that has received increased attention in the research on ratings is Lawler's (1967) multitrait-multirater approach, which is a special case of the multitrait-multimethod approach (Campbell and Fiske, 1959). While the number of researchers using this approach is still relatively small, the approach became more attractive when it was conceptually linked (Kavanagh, MacKinney, and Wolins, 1971) with the analysis of variance approach proposed by Guilford (1954).

State of the Art

A review of recent studies that have dealt either directly or indirectly with rating errors or other indices of rating quality confirmed the fact that confusion and inconsistency are still very much with us. Specifically, an examination of approximately thirty papers published in Journal of Applied Psychology, Organizational Behavior and Human Performance and Personnel Psychology during the years 1974-77 revealed only too clearly the ambiguity and lack of consistency apparent in our review of textbooks. Additionally, there is less than unanimous agreement regarding the operational definitions of several other criteria of rating quality, including discrimination among ratees and interrater reliability or agreement.

Leniency/severity has been defined as (1) a rating level effect, (2) a tendency to assign a higher/lower rating to an individual than is warranted by that individual's performance, (3) a response set attributed to easy/hard-nosed raters whose ratings are consistently higher/lower than warranted with respect to an external criterion of known true performance, and (4) a shift in mean ratings from the midpoint of the scale in the favorable/unfavorable direction. The confusion is obvious. The second and third definitions refer explicitly or implicitly to a "true" performance score -- an "eye-of-God-reality" (Lumsden, 1976) to which mortal psychologists must remain forever blind. The fourth definition suggests that average performance levels should coincide with the midpoints of rating scales -- an assumption employers strive to invalidate with effective personnel selection and training programs. We are therefore left with leniency/severity as a rating level effect -- a rather shallow and unsatisfying definition.

Three operational definitions of leniency/severity emerged in the literature reviewed. By far the most common strategy entailed a simple examination of mean dimension ratings, as compared with the scale midpoint. Much less popular approaches were looking at the degree of skewness that characterized distributions of dimension ratings, and searching for a significant Rater main effect in a Rater X Ratee X Dimension analysis of variance.

Halo has been described as (1) a tendency to attend to a global impression or each ratee, rather than carefully distinguishing among levels of performance exhibited by individual ratees on different performance dimensions; (2) a tendency to place a given ratee at the same level on different dimensions; and (3) an inability or unwillingness to discriminate among different aspects of a ratee's performance. Although there seems to be less confusion here than in the case of leniency/severity, a certain amount of imprecision remains.

Perhaps this imprecision is best described by Nisbett and Wilson (1977), who distinguished between a "strong" interpretation of the halo effect and a weaker interpretation. Briefly stated, a strong interpretation implies that global evaluations actually alter evaluations of more specific attributes about which the rater has information fully sufficient to permit independent assessment. A weaker interpretation implies merely that a global evaluation "might color presumptions about specific traits or influence interpretation of the meaning or affective value of ambiguous trait information." Nisbett and Wilson further stated that an experimental approach is necessary in order to demonstrate the strong interpretation of the halo effect.

The literature review revealed four operational definitions of halo. One strategy examines the intercorrelations among dimensions ratings. A closely related strategy is based upon a factor analysis of the dimension intercorrelation matrix -- a greater number of emergent interpretable factors (or components) suggests less halo. A third approach focuses on the variance (or standard deviation) associated with a particular rater's ratings of individual ratees across the performance dimensions -- greater dispersion of ratings across dimensions suggests less halo. A fourth strategy looks for a significant Rater X Ratee interaction in a Rater X Ratee X Dimension analysis of variance.

Central tendency, a third rating error commonly described in industrial psychology texts, was almost ignored in the literature reviewed. Colloquially defined as a rater's unwillingness to go out on a proverbial limb in either direction (De Cotiis, 1977), operational definitions have focused on (1) the proximity of mean dimension ratings (as well as the magnitude of their associated standard deviations) to the (neutral) midpoint of the rating scale, and (2) the degree of leptokurtosis that characterizes distributions of dimension ratings.

Although the term central tendency was generally avoided, many researchers examined ratee discrimination as an index of the quality of performance ratings. This is often described negatively as the problem of restriction of range. The most common operational definition involves the standard deviations associated with mean dimension ratings across all ratees -- larger standard deviations (or variances) reflect greater use of the entire rating scale on a given dimension, and therefore less range restriction and greater discrimination among ratees. A second strategy searches for a significant Ratee main effect, and/or a significant Ratee X Dimension interaction in a Rater X Ratee X Dimension analysis of variance. The concept of discriminant validity and Kavanagh, MacKinney, & Wolins' (1971) analysis of variance technique for interpreting multitrait-multimethod/multirater data have also been discussed in the context of this index of performance rating quality.

Another quality index that is conceptually and analytically similar to central tendency is reliability. Since reliability requires ratee variance the lack of this variance would result in the lack of reliability. Although several authors have expressed some reservations regarding the status of interrater reliability or agreement as a criterion of rating quality, it has commonly been invoked as such in the literature reviewed. Published indices of interrater reliability or agreement have included (1) the standard deviations of the ratings assigned to a given ratee on the various dimensions; (2) correlations for pairs of raters evaluating the same ratee on identical performance dimensions (similar to the concept of convergent validity in the context of multitrait-multimethod/multirater matrices); (3) intraclass correlations; (4) ratee main effects in Rater X Ratee X Dimension analyses of variance; and (5) the absence of significant Rater X Ratee interactions in Rater X Ratee X Dimension analyses of variance. Additionally, Schmidt and Hunter (1977) have suggested that the appropriate

reliability coefficient for ratings used as criteria is the correlation across a reasonable time interval between ratings produced by different raters at Time 1 and Time 2. Once again, confusion (or worse, according to Lumsden, 1976) appears to reign.

Other criteria for judging the adequacy of performance ratings have included the extent to which similarity error, contrast error, and first impressions error are minimized, as well as various rater reactions during the rating process (preference for a given rating scale or format, relevance, difficulty, perceived accuracy of ratings, etc.). Since our concern is primarily psychometric, these criteria are not examined in detail here.

The preceding review amply demonstrates the disarray of the field, and points to the need for a reconceptualization of the problem and an integration of the various approaches. A second major problem encountered was a failure to clearly differentiate between the rater's behavior and the consequences of that behavior when applied over different raters, ratees, dimensions, and situations. Definitions were generally stated in terms of an individual rater's behavior, but most indices of the errors are expressed in terms of aggregates of raters and ratees, and over various dimensions and sometimes situations (or methods).

A third, and more difficult problem associated with research focusing on ratings has been the use of several types of designs or data collection procedures. These different designs have inherently limited capabilities for aggregating data, and while apparently similar indices can be derived, the indices reflect the limitations. The ideal data design is a full Rater by Ratee by Dimension situation, with all raters rating all ratees on all dimensions. This design, while ideal, is seldom achieved. The most common design is a partial Rater by Ratee by Dimensions design, where some raters rate some ratees on all dimensions. Often this design is reduced into a Ratee by Dimensions design summed (or

averaged) over raters. A third design is the Ratee by Dimensions design with a single rater per ratee (may include multiple ratees per rater). A fourth design is a Rater by Dimension design with only one (or sometimes more) ratees.

The fourth and final issue is the purpose of the analysis. While much of the original work in rating was aimed at assessing the absolute amount of rating error in a particular rating system, more recent research efforts have been directed at the relative amounts of error inherent in two (or more) rating systems. (This second approach has introduced the use of dependent variables which are assumed to directly reflect error, rather than a summary score which can be used to evaluate error.) For example, some researcher (e.g., Bernardin & Walter, 1977) have used each rater's dispersion in ratings as a score in analyzing rater differences.

Where Do We Go From Here?

The approach outlined below is designed to dea; with each of four issues raised in this paper by: (1) suggesting standard definitions, (2) couching definitions in terms of aggregate ratee/rater/dimension data rather than rater behavior; (3) applying definitions across several different data collection strategies or research designs and (4) specifying dependent measures.

One approach investigated and rejected, originally suggested by Guilford (1954), emphasized an analysis or variance approach. While Hoyt (1941) had previously used an ANOV approach to estimate reliability, Guilford identified some of the main and interaction effects in the full three way design as specific rating errors. Willingham and Jones (1958) expanded on Guilford's concepts (also see Stanley, 1961, and Maxwell & Pilliner, 1968) by labeling all main and interaction effects as types of "errors". A separate approach suggested by Kavanagh, MacKinney and Wolins (1971) focused on a multitrait-multirater matrix (Lawler, 1967). Several problems with the rater by ratee by dimension approach

must be acknowledged. First, the use of dimensions as a repeated factor presents some conceptual as well as analytical difficulties. Conceptually, a repeated factor represents repeated observations of the same subject using the same dependent measure under several different conditions or over several different occasions. The extent to which ratings over several dimensions constitute the same dependent measure (the same anchors notwithstanding) is highly questionable; e.g., height and weight are both related concepts and are indicative of size but are not the same dependent measure. Analytically, we require correlations between repeated measures (i.e., dimensions), but conceptually we desire low or no correlations. To the extent that we achieve the absence of any correlations among dimensions, we reduce the appropriateness of a repeated measures design. For these reasons we rejected the rater by ratee by dimensions approach.

Many of the confusions associated with rating errors may well be a function of negatively toned definitions, and much might be gained by defining and emphasizing desirable characteristics in a rating system. We submit that desirable characteristics of a rating system include the following:

1. Ratee Differences - Rating systems should result in the ability to adequately distinguish between and among ratees. The use of this positive statement clearly suggests the relationship of this characteristic to the traditional notion of interrater reliability, an index that reflects the ratio of summed rater variance (over ratees) to ratee dispersion. This property closely approximates, or is analogous to central tendency.
2. Rater Comparability - The averages of raters' ratings should be comparable when their ratings are based on comparable ratees. This requirement becomes increasingly important as the number of raters

evaluating a single ratee drops, and the impact of any one rater on that ratee's evaluation therefore increases. This is analogous to the absence of leniency or severity error, although the idea of interrater agreement is also suggested.

3. Ratees should be rank-ordered differently on separate dimensions -- Ratees should generally receive different numerical ratings on different dimensions. Note that there is no requirement for a complete lack of relationship among dimensions. This is analogous to the absence of halo error.

The first two characteristics are defined in terms of a Rater X Ratee matrix and may be defined as Ratee main effect and the absence of a Rater main effect, respectively in an ANOV. (Note that some indices of interrater reliability do not demand rater comparability, since reliability here depends upon rater's ratings being correlated, irrespective of the absolute levels of assigned ratings. Reliability does depend upon ratee differences, however.) Each dimension requires a separate ANOV. The third suggested characteristic is analogous to hypothesizing that a ratee by dimension matrix (summed over raters) involves more unique dimension variance than common variance across dimensions or simply that dimensions are not correlated. This characteristic can be evaluated with a number of correlational strategies including a principal components analysis or a simple average correlation between a given dimension and all other dimensions.

The characteristics that can be evaluated with each of the data collection procedures described above are as follows:

1. Rater X Ratee X Dimension: Characteristics 1, 2 and 3
2. Rater X Ratee : Characteristics 1 and 2
3. Ratee X Dimension : Characteristic 3
   (but based only on a single rater)

4. Ratee X Dimension (summed across raters) : Characteristic 3 (Note: this occurs when some of the raters evaluate some of the ratees.)

5. Ratee X Dimension : None

Of course, generalizations that emerge from any of these designs are restricted by situational parameters.

The restrictions imposed by the "reduced" data collection procedures severely limit the usefulness of strategies 3 and 5. Limiting the collection of data to one rater (strategy 3) does not permit any distinction between problems associated with the rating procedure and problems associated with the rater. Limiting the study to one ratee (strategy 5) suffers from the same problem of differentiating between ratee and rating system difficulties. Using an analogy from the testing area, it would be similar to evaluating the differences between "multiple choice" and "fill in the blank" methods using a one item test or evaluating these same methods with many items using only one person (respondent) -- in neither case would the results be considered satisfactory.

One further issue that needs to be explored is a strategy for comparing the relative effectiveness of two different rating procedures. An ANOV approach, where method (or condition) is introduced as an additional factor in the analysis could be used. A related problem is the extent to which different types of raters (peers, self, supervisors) may produce ratings of different quality. The multitrait-multimethod matrix approach might be appropriate here.

Summary and Conclusions

In this paper we have briefly traced the development of some psychometric indices of rating quality, and have described in considerable detail the conceptual and analytical confusion that characterizes contemporary research in this area. A set of positively toned criteria for evaluating the quality of rating data was suggested, and the suitability of several common research designs for generating

statistical indices of those criteria was examined. Given the potentially large impact of ratings on many aspects of our lives, the need for some sort of quality control over such ratings seems obvious. Such control is unlikely to be realized, however, unless the behavioral scientists who study rating behavior learn to speak the same language and communicate meaningfully with each other. It is hoped that the conceptual and analytical suggestions proffered here will facilitate the adoption of some such common vocabulary, and thereby contribute to a more profound understanding of the psychosomatic pitfalls inherent in asking one human being to evaluate one or more traits, characteristics or behaviors of another.

References

Barrett, R. S. Performance ratings. Science Research Associates, 1966.

Bernardin, H. J., & Walter, C. S. Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 1977, 62, 64-69.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Cummings, L. L., & Schwab, D. P. Performance in organizations: determinants and appraisal. Glenview, Il.: Scott, Foreman & Co., 1973.

De Cotiis, T. A. An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 1977, 19, 247-266.

Dunnette, M. D. Personnel selection and placement. Belmont, Ca.: Wadsworth Publishing Co., Inc., 1966.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Ford, A. Neutralizing inequalities in ratings. Personnel Journal, 1931, 9, 466-469.

Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.

Hays, W. L. Statistics. New York: Holt, Rinehart and Winston, 1963.

Hoyt, C. J. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Johnson, D. M. Reanalysis of experimental halo effects. Journal of Applied Psychology, 1963, 47, 46-47.

Johnson, D. M., & Vidulich, R. N. Experimental manipulation of the halo effect. Journal of Applied Psychology, 1956, 40, 130-134.

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. Issues in managerial performance: multitrait-multimethod analysis of ratings. Psychological Bulletin, 1971, 75, 34-49.

Kingsbury, F. A. Analyzing ratings and training raters. Journal of Personnel Research, 1922, 8-9, 377-383.

Kneeland, N. These lenient tendencies in rating. Personnel Journal, 1929, 7, 356-366.

Landy, F. J., & Trumbo, D. A. Psychology of Work Behavior. Homewood, Il.: Dorsey, 1976.

Lawler, E. E. The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, 51, 369-381.

Lunsden, J. Test theory. In M. R. Rosenzweig & L. W. Porter (Eds.), Annual Review of Psychology, (Vol 27). Palo Alto, Ca.: Annual Review Inc., 1976.

Maxwell, A. E., & Pilliner, A. E. G. Deriving coefficients of reliability and agreement for ratings. British Journal of Mathematical and Statistical Psychology, 1968, 21, 105-116.

McCormick, E. J., & Tiffin, J. Industrial Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1974.

Murray, H. A. Explorations in personality. New York: Oxford University Press, 1938.

Newcomb, T. An experiment designed to test the validity of a rating technique. Journal of Educational Psychology, 1931, 22, 279-289.

Nisbett, R. E., & DeCamp, W. T. The halo effect: Evidence for unconscious alteration of judgments. Journal of Personality and Social Psychology, 1977, 35, 250-256.

Rugg, H. Is the rating of human character practicable? Journal of Educational Psychology, 1921, 12, 485-501.a

Rugg, H. Is the rating of human character practicable? Journal of Educational Psychology, 1921, 12, 485-501.b

Rugg, H. Is the rating of human character practicable? Journal of Educational Psychology, 1922, 13, 30-42.a

Rugg, H. Is the rating of human character practicable? Journal of Educational Psychology, 1922, 13, 81-93.b

Schmidt, F. L., & Hunter, J. E. Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 1977, 62, 529-540.

Schneider, D. J. Implicit personality theory: A review. Psychological Bulletin, 1973, 79, 294-309.

Schultz, D. P. Psychology and Industry Today. New York: MacMillan Publishing Company, Inc., 1978.

Stanley, J. C. Analysis of unreplicated three-way classification with applications to rater bias and trait independence. Psychometrika, 1961, 26, 205-219.

Stockford, L., & Bissel, H. W. Factors involved in establishing a merit-rating scale. Personnel, 1949, 26, 94-116.

Smith, P. C. Behaviors, results, and organizational effectiveness: The problem of criteria. In M. D. Dunnette (Ed.), Handbook of Industrial and Organizational Psychology. Chicago: Rand McNally College Publishing Co., 1976.

Thorndike, E. L. A constant error in psychological ratings. Journal of Applied Psychology, 1920, 4, 25-29.

Webb, E. Character and intelligence. British Journal of Psychology, 1915, 3, 1-99.

Wells, F. L. A statistical study of literary merit. Archives of psychology, 1909, 7.

Willingham, W. W., & Jones, M. B. On the identification of halo through ANOV. Educational and Psychological Measurement. 1958, 18, 403-407.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.

Table 1 REVIEW OF DEFINITIONS AND QUANTIFICATIONS OF RATING ERRORS

Source		Type of Rater Error: Leniency	Central Tendency	Halo
Guilford (1954)	D[a]	Rates too high/low	Hesitates to give extreme judgments	Judge in terms of a general mental attitude
	Q[b]	Differences in rater means	Reduced ratee variance	Trait correlations
Barrett (1966)	D	(In terms of quantification)	Rates in scale center when not clear as to the meaning	Judge in terms of a general mental attitude
	Q	Grand mean above scale mid-point	None given	Trait correlations
Dunnette (1966)	D	Says only good things	Failing to discriminate	Describe with all good or bad sounding statements
	Q	None given	None given	None given
Cummings and Schwab (1973)	D	Rates too high/low	Give everyone average ratings	Judge in terms of only one characteristic
	Q	None given	None given	Trait correlations (implied)
McCormick and Tiffin (1974)	D	Ratings concentrated high (low)	Ratings concentrated in the middle	Judge in terms of only one characteristic
	Q	Differences in rater means	Differences in rater means	Trait correlations
Landy and Trumbo (1976)	D	Unusually hard or easy	Unwillingness to assign extreme scores	Generally favorable or unfavorable impression
	Q	None given	Reduced ratee variance	None given
Smith (1976)	D	(In terms of quantification)	(In terms of quantification)	Ratings affect other ratings
	Q	Displaced mean (from mid-point) skewness	Leptokurtic distribution	Trait correlations

[a]Definition
[b]Quantification